

## FEATURE ARTICLE

### Determining the Role of Hydration Forces in Protein Folding

**Jon M. Sorenson**

*Department of Chemistry, University of California, Berkeley, Berkeley, California 94720*

**Greg Hura**

*Graduate Group in Biophysics, University of California, Berkeley and Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720*

**Alan K. Soper**

*ISIS Facility, Rutherford Appleton Laboratory, Chilton, Didcot, Oxon OX11 0QX, U.K.*

**Alexander Pertsemlidis**

*Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75235-9038*

**Teresa Head-Gordon\***

*Life Sciences Division & Physical Biosciences Division, Lawrence Berkeley, National Laboratory, Berkeley, California 94720*

*Received: February 5, 1999; In Final Form: April 19, 1999*

This article describes a combined experimental, theoretical, and computational effort to show how the complexity of aqueous hydration can influence the structure, folding and aggregation, and stability of model protein systems. The unification of the theoretical and experimental work is the development or discovery of effective amino acid interactions that implicitly include the effects of aqueous solvent. We show that consideration of the full range of complexity of aqueous hydration forces such as many-body effects, long-ranged character of aqueous solvation, and the assumptions made about the degree of protein hydrophobicity can directly impact the observed structure, folding, and stability of model protein systems.

#### Introduction

One of the primary issues in protein folding is determining what forces drive folding and eventually stabilize the native state.<sup>1</sup> A delicate balance exists between electrostatic forces such as hydrogen bonding and salt bridges, and the hydrophobic effect, which are present for both intramolecular protein

interactions and intermolecular contributions with the surrounding aqueous environment. An additional layer of complexity arises from how these forces are modulated by the varied chemical properties of the individual amino acids, whose local conformations and energetics are influenced by the intrinsic secondary structure propensities and/or the primary sequence context within which the residues reside.

Energy landscape models have defined a “new” view of

\* To whom correspondence should be addressed.

protein folding for explaining the kinetics and thermodynamics of folding.<sup>2–5</sup> This recent theoretical work in protein folding has suggested a free energy surface that is funnel-like in shape; i.e., as folding progresses, the energy decreases faster than the entropy as measured by the reduction in the number of states. The folding surface is also characterized by a significant energy gap separating the native state from the nearest, structurally dissimilar non-native state. Given the growing consensus on the validity of energy landscape models, more recent work has focused on experimental confirmation of predictions made about folding intermediates<sup>3</sup> and should in the future include a better understanding as to what are the specifics for shaping a good folding free energy landscape.

There have been many suggestions for the protein folding mechanism and intermediates from both experiment and theory that attempt to define specific attributes such as explicit sequence biochemistry as well as primary physical forces.<sup>1,6–32</sup> The framework model,<sup>17</sup> for example, emphasizes the biochemical aspects of the sequence and the secondary structure propensities of amino acids as contributing to the earliest biases in the protein folding pathway, and the conformational search space is then narrowed to efficiently find the native structure. The framework model is motivated by the experimental observation of early-formed secondary structure in trapped kinetic intermediates of ribonuclease A and cytochrome C found by hydrogen-exchange labeling and proton NMR.<sup>15–17</sup> On the other extreme, models that reduce amino acid individuality into hydrophobic and polar “flavors” place more emphasis on physical forces such as the hydrophobic effect or hydrogen bonding. For example, one physical-based model describes the earliest folding events as being dominated by the thermodynamics associated with hydrophobic interactions, biases that result in a collapse of the polypeptide chain to a compact state.<sup>8</sup> Some folding experiments have determined the “collapse” mechanism to be operative for small proteins, while the framework model is most consistent for larger proteins.<sup>24,25</sup>

The complexity of the protein folding problem has led to consideration of simplified representations that might serve as reduced models of the protein folding process. There has been considerable research devoted to understanding exclusively hydrophobic phenomena<sup>19–23</sup> and electrostatic contributions,<sup>26,27</sup> and the consequences each has on protein folding and stability. Much work has been devoted to understanding the formation of secondary structure elements<sup>28–32</sup> such as turns, helices, and sheets for small peptides and has been very instructive in indicating under what conditions secondary structures of small peptide fragments can serve as folding initiates.

Our group’s effort over the past several years has been directed toward a model systems approach for characterizing hydration forces between amino acid solutes that are relevant in the context of protein folding.<sup>33–39</sup> We have used both theoretical and experimental approaches: molecular dynamics simulations,<sup>33,35–38</sup> neutron and X-ray solution scattering experiments,<sup>34–38</sup> and protein folding models.<sup>39</sup> The unification of the theoretical and experimental work is the development or discovery of potentials of mean force between amino acids that implicitly include different aspects of aqueous hydration forces, such as many-body effects or more long-ranged character of aqueous solvation.

The theoretical conclusions made from energy landscape views are largely based on highly idealized lattice models of proteins that have no atomic detail and use very nonspecific descriptions of residue–residue interactions.<sup>2,40–42</sup> We have attempted to place more physical emphasis on how residue

interactions would give rise to a funneled landscape by investigating the effect of adding features of hydration forces to simple protein folding lattice models.<sup>39</sup> This protein folding study investigated the effect of adding a multibody description of hydration to a simple two-flavor lattice protein model.<sup>39</sup> Sequences in the hydrated model were more frequently found to have unique ground states, to fold faster, and to fold with more cooperativity than sequences in the corresponding model without solvation terms. Our results indicate that the introduction of physically motivated solvation terms can improve the poor performance of two-flavor lattice models, since the multibodied nature of hydration mimics amino acid diversity, which in turn gives rise to a more cooperative folding transition.<sup>39</sup>

The demonstration that model hydration forces can alter the kinetics and/or thermodynamics of protein folding models provides important interplay to our solution scattering experiments and simulations that attempt to determine hydration forces from simulation and experiment. We describe our studies of hydration forces for dilute concentrations of amino acids, with characterization of the corresponding changes in water structure, and illustrate this for solutions of a common hydrophobic amino acid, *N*-acetyl-leucine-amide (NALA).<sup>34–37</sup> These solution scattering studies constitute a model of the solvation structure and free energy of amino acid association during early protein folding events. By combining information from solution scattering experiments with molecular dynamics simulation, we demonstrate that important information in the small-angle scattering region of these experiments can be mined to resolve solute–solute correlations, their length scales, and thermodynamic consequences, even at dilute concentrations.<sup>37</sup>

Solution scattering experiments and simulations can also be used to probe solvation for more concentrated aqueous solutions of hydrophobic solutes and suggest a model of later protein folding events when significant spatial domains of the protein comprise a hydrophobic core. We describe preliminary X-ray solution scattering results on the behavior of the hydrophobic amino acid, *N*-acetyl-leucine-methylamide (NALMA) in water as the concentration of the amino acid increases.<sup>38</sup> Our experimental and simulation results suggest that later protein folding events would involve both monodispersed amino acids and the formation of small clusters (between two and six) of hydrophobic residues. The experimental data over the full range of concentration, interpreted by molecular dynamics simulation of the same X-ray experiments, appear to be inconsistent with the hydrophobic solutes segregating themselves completely from the aqueous solvent to form a large hydrophobic cluster.

### Hydration Forces as Biases in Protein Folding Free Energy Landscapes

Nearly all lattice and many off-lattice studies designed to investigate protein folding do not include explicit residue–water and water–water interactions, and any implicit hydration contained in typical lattice model parameters ignores two prominent features of hydration forces: their many-body nature and their potentially long-range effects. One aim of our recent work has been to examine the effect of adding a simple multibody potential on the conclusions drawn previously from studies of lattice models with pairwise-additive energies.<sup>39</sup>

For our lattice studies we simulated 36 residue chains as self-avoiding walks on a cubic lattice with each residue represented by a single interaction site. Details of the lattice model simulation protocol are discussed elsewhere.<sup>39</sup> The energy for a typical lattice folding study involves an energy function of the form

$$E = \sum_{i < j}^N B_{ij} \Delta_{ij} \quad (1)$$

where the double sum is over the  $N$  residues of the chain,  $B_{ij}$  is the contact energy between residues  $i$  and  $j$ , and  $\Delta_{ij}$  is 1 if residues  $i$  and  $j$  are nearest neighbors and not contiguous on the chain, and 0 otherwise. A solvation model for lattice folding studies was designed that captures several aspects of hydration: different free energies of solvation for hydrophobic and polar residues, multibody effects, and long-range effects. We accomplish this by redefining the contact energy matrix elements to be

$$B_{ij} = (1 - \lambda_{ij})B_{ij}^u + \lambda_{ij}B_{ij}^f \quad (2)$$

where  $B_{ij}^u$  represents the contact energy matrix element for the unfolded chain,  $B_{ij}^f$  is the contact energy matrix element for the folded chain, and  $\lambda_{ij}$  is a bond solvation parameter representing the degree of solvation of the  $ij$ th contact. Our current study is a two-flavor model in which the type of each residue is restricted to be either hydrophobic (H) or polar (P). For the unfolded chain contact energy matrix we chose

$$\mathbf{B}^u = \begin{matrix} & \begin{matrix} \text{H} & \text{P} \end{matrix} \\ \begin{matrix} \text{H} \\ \text{P} \end{matrix} & \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \end{matrix} \quad (3)$$

and for the folded matrix

$$\mathbf{B}^f = \begin{matrix} & \begin{matrix} \text{H} & \text{P} \end{matrix} \\ \begin{matrix} \text{H} \\ \text{P} \end{matrix} & \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \end{matrix} \quad (4)$$

The form of the unfolded matrix is motivated by our own experimental and simulation work,<sup>34–37</sup> described in the next sections, which has been focused on probing the length scales over which hydrophobic amino acids are attracted to each other in water and whether interactions between hydrophilic amino acids in water are repulsive. The folded matrix in eq 4 is similar to a form studied in previous theoretical,<sup>43,44</sup> design,<sup>45,46</sup> and simulation<sup>47,48</sup> studies but differs from this previous work in that the average interaction energy is more repulsive. As a comparison, we also performed simulations in the nonsolvation model using the folding matrix alone, i.e., with  $B_{ij} = B_{ij}^f$  in eq 2.

We let the energy of contacts interpolate between a matrix of unfolded contact energies and a matrix of folded contact energies, with  $0 \leq \lambda_{ij} \leq 1$ . The interpolation parameter,  $\lambda_{ij}$ , represents the degree to which a particular contact is solvated and has the following form:

$$\lambda_{ij} = \frac{\lambda_i + \lambda_j}{2} \quad (5)$$

where the individual monomer solvation parameter  $\lambda_i$  is defined as

$$\lambda_i = s_i/s_i^0 \quad s_i/s_i^0 < 1 \\ = 1 \quad \text{otherwise} \quad (6)$$

$s_i$  is a measure of the solvent-accessible surface area of monomer  $i$ ,

$$s_i = \sum_j^N \Delta_{ij} \quad (7)$$

**TABLE 1: Representation of the Solvation Model as a Multiflavor Model<sup>a</sup>**

	H0	H1	H2	H3	P0	P1	P2
H0	−1	−1	−1	−1	0	0.25	0.50
H1	−1	−1	−1	−1	0.167	0.417	0.667
H2	−1	−1	−1	−1	0.333	0.583	0.833
H3	−1	−1	−1	−1	0.50	0.75	1
P0	0	0.167	0.333	0.50	1	0.50	0.0
P1	0.25	0.417	0.583	0.75	0.50	0	−0.50
P2	0.50	0.667	0.833	1	0	−0.50	−1

<sup>a</sup> The number after the residue type is the solvation state  $s_i$  (eq 7). Flavors H4, H5, P3, P4, and P5 are not shown because by eq 6 they are equivalent to flavors with lower solvation states. H0 and P0 do not actually occur in simulation because energies are only present between residues in contact, and the presence of a single contact would necessarily raise the solvation state above zero.

**TABLE 2: Properties of the Foldable Sequences Studied with the Solvation and Nonsolvation Models<sup>a</sup>**

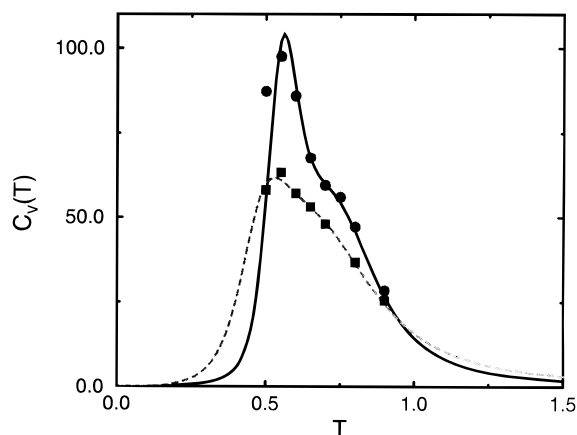
sequence	solvation				nonsolvation			
	$E_{\min}$	$\tau_{\text{MFPT}} \times 10^8$	$T_f$	$T_f/T_g$	$E_{\min}$	$\tau_{\text{MFPT}} \times 10^8$	$T_f$	$T_f/T_g$
1	−35.17	5.2(8)	0.58	1.15	−36.00	6.2(8)	0.48	0.87
3	−34.50	1.3(2)	0.57	1.09	−34.00	6.9(6)	0.34	0.67
6	−36.00	1.0(1)	0.64	1.30	−36.00	1.0(3)	0.50	0.99
20	−36.50	0.9(2)	0.64	1.23	−36.00			
26	−35.92	0.5(1)	0.64	1.30	−36.00			
29	−35.83	1.0(1)	0.54	1.02	−36.00	1.2(2)	0.40	0.79
30	−36.00	1.5(4)	0.55	1.00	−36.00			
35	−35.17	1.9(3)	0.56	1.05	−36.00			

<sup>a</sup> The uncertainty in the last digit is given in parentheses.  $E_{\min}$  is the native state energy,  $\tau_{\text{MFPT}}$  is the mean first-passage time,  $T_g$  is the kinetic glass temperature,  $T_f$  is the folding temperature found from the tangent construction with the density of states.<sup>51</sup> All quantities in reduced units.

and  $s_i^0$  is a measure of the optimal solvation state for residue  $i$ . We chose  $s_i^0 = 2$  for polar residues and  $s_i^0 = 3$  for hydrophobic residues to represent the tendency for hydrophobic residues to bury themselves in the protein interior, away from solvent.

Using standard sequence design methods,<sup>39,45</sup> we found eight foldable sequences for study with the solvation model. To validate studying the same sequence in both the solvation and nonsolvation models, we verified that the sequences studied with and without solvation were optimally designed sequences. Four of the eight sequences were found to have degenerate ground states without solvation and consequently could not be used for nonsolvation folding studies. That only four of the eight foldable sequences had nondegenerate native states in the nonsolvation model indicates that incorporation of multibody solvation lifts the degeneracy that has been observed for two-flavor lattice models.<sup>49,50</sup> This observation is not surprising when we recast our model as a multiflavor model, since multiflavor models in general have more sequences with nondegenerate ground states.<sup>41,50</sup> Table 1 shows the contact energies when the solvation model is reformulated as a multiflavor model. The difference between our solvation model and a true multiflavor model is that the flavors of each monomer in our model are environment-dependent and are able to change over the course of the simulation. In essence, the protein sequence is given some freedom to redesign itself as it folds.

Table 2 compares various folding properties of these sequences and properties of their native structures for the solvation and nonsolvation models. The folding kinetics were explored for each sequence by varying the temperature and collecting statistics on mean first-passage times for folding to a collapsed state ( $\geq 36$  contacts), folding to a compact state (40



**Figure 1.** Heat capacity ( $C_v$ ) vs temperature for sequence 6 in the solvation (solid line) and nonsolvation models (dashed line). The curves were generated with the histogram Monte Carlo method; the points are taken from Monte Carlo simulations at those temperatures.

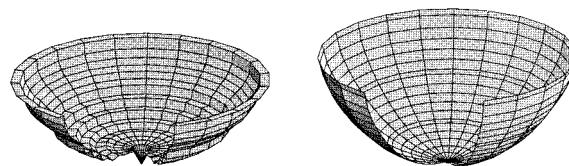
contacts), and folding to the native state. If in a particular run a sequence was found not to fold within the maximum simulation time of  $10^9$  steps, we averaged the maximum time into the mean.<sup>48</sup> As such, the reported times are all lower bounds to the true mean first-passage times. Each sequence folds faster under the solvation model, although the extent of this varies (Table 2).

We note that there are several possible definitions of the folding temperature  $T_f$ . We prefer a definition of the folding temperature in which the free energy of the native state is equal to the free energy minimum of the unfolded states.<sup>51–53</sup> The histogram Monte Carlo method<sup>47,54</sup> allows us to determine  $T_f$  in this way, or equivalently, by a tangent construction using the density of states  $\Omega(E)$ .<sup>51</sup> The accuracy of the calculated  $\Omega(E)$  was confirmed by calculating curves for  $E$  vs  $T$  and  $C_v$  vs  $T$  and comparing these curves to that found by simple averaging from Monte Carlo simulations at various temperatures (Figure 1). Table 2 shows that the folding temperature is consistently higher for the sequences under the solvation model than with the nonsolvation model.

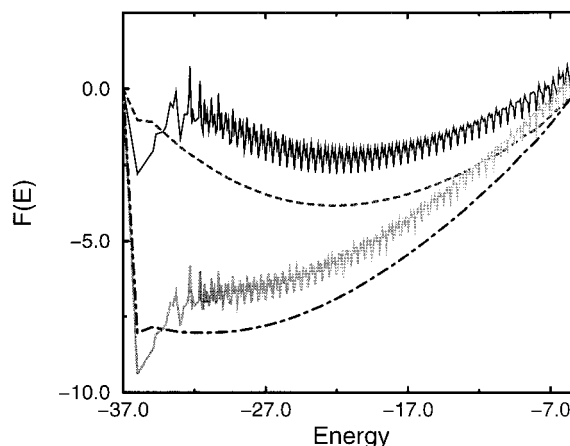
We define the glass transition temperature,  $T_g$ , as the temperature at which the folding time is halfway between the maximum simulation time,  $\tau_{\max}$ , and the fastest folding time for that sequence.<sup>48</sup> Good folding sequences should have folding temperatures above the glass temperature.<sup>48</sup>

Of particular interest for comparing minimalist protein folding models with experiment is the ratio of the folding temperature to the glass transition temperature,  $T_f/T_g$ .<sup>55</sup> This ratio gives a simple characterization of the steepness of the protein folding funnel for theoretical and real proteins. It has been found that  $T_f/T_g$  can be as high as 1.3 for two-flavor models, while the ratio for a real protein is predicted to be approximately 1.6.<sup>55,56</sup> Lattice models with more flavors have been found to have a  $T_f/T_g$  ratio closer to that for real proteins.<sup>55,57</sup>

All eight sequences in the solvation model are good folders based on the  $T_f/T_g$  ratio, while the four sequences in the nonsolvation model are bad folders using this criteria. We note that the  $T_f/T_g$  ratios for our solvation model, while much better than for the nonsolvation model, are not exceptional in an absolute sense. One possible reason is that we did not optimize the matrices of interactions (eqs 3 and 4), and we would anticipate better optimized interactions to produce better ratios. Second,  $T_g$  has some dependence on the definition of  $\tau_{\max}$ , which is  $10^9$  steps in our study. This is especially a concern for calculating  $T_g$  for the slower folding sequences in the nonsol-



**Figure 2.** Free energy vs entropy for sequence 6 in the solvation (left) and nonsolvation models (right) at  $T/T_f = 0.78$ . The depth of the funnel corresponds to the free energy, and the radial coordinate is the entropy.



**Figure 3.** Free energy vs energy for sequence 6 in the solvation model and nonsolvation model:  $T = 0.636$ , solvation model (solid line);  $T = 0.636$ , nonsolvation model (dashed line);  $T = 0.487$ , solvation model (dotted line);  $T = 0.487$ , nonsolvation model (dotted-dashed line). The scale for the free energies is a relative scale; for comparison, the curves shown here were offset to make  $F(-5) = 0$ .

vation model, and even the folding times of our fastest sequences are only an order of magnitude less than  $\tau_{\max}$ . The extent of the problem was tested for sequence 6 by running simulations at low temperatures for  $\tau_{\max} = 10^{10}$  steps; the resulting prediction for the kinetic  $T_g$  was shifted to lower temperature by  $\sim 10\%$  and therefore increased our  $T_f/T_g$  ratios by about 10%. Nonetheless, the higher  $T_f/T_g$  ratio for sequences under the solvation model show that the addition of solvation has shaped a better folding free energy surface.

These combined results indicate that the addition of solvation terms to a two-flavor model changes the underlying free energy landscape, as shown in Figure 2. At  $T/T_f = 0.78$ , the native state of sequence 6 is favorable enough to create a marked depression at the center of the funnel for the solvation model. At the corresponding temperature for the same sequence in the nonsolvation model the native state is not as stable; i.e., we see much less bias pulling the center of the funnel down, and there is a multitude of competing collapsed states with very similar free energies. From Figure 1 we can see that the heat capacity curve is much sharper and more peaked for the solvation model, characteristic of a more first-order-like transition.

While the folding of sequence 6 in the solvation model is more cooperative, the sequence folds at comparable speeds in both models. The origin of the observed kinetics becomes clearer when we examine the free energy of folding versus energy,  $F(E)$  at the temperatures where the free energy of the folded state equals that of the minimum free energy of the unfolded states (Figure 3). The curves in Figure 3 for the nonsolvation model barely exhibit two minima, a prerequisite for two-state kinetics, while the solvation model curves exhibit features that support good two-state kinetics. However, the rougher free energy curve in the solvation model works against this sequence, with many stable traps in the region of the transition ensemble that hamper fast folding (Figure 2). The slow steps of folding are searching



through the plateau of partially collapsed states just above the native state in free energy, an entropic bottleneck in the energy landscape that has been described as a champagne glass landscape.<sup>5</sup>

### Is There a Causal Effect between Hydration Forces and Altered Water Structure?

Long-ranged hydration forces between *macroscopic surfaces* are well-established in the literature.<sup>58–67</sup> For these extended surfaces, the measured forces are far greater than those based on electrostatic interactions, steric repulsion, and van der Waals forces, with measurable effects often extending to distances greater than 10 Å.<sup>60–62</sup> However, various types of collective processes likely underlie the long-ranged interactions for extended surfaces,<sup>67,68</sup> such as a dewetting transition,<sup>20,21</sup> that would likely not apply to single amino acid solutes or finite length polypeptide chains in water. What is applicable to smaller solutes, or atomistic views of polypeptide chains, is that hydration waters near hydrophobic groups are more ordered than bulk water. One hypothesis we have explored is that alterations in water structure around amino acids give rise to hydration forces that correlate amino acids over longer length scales than, for example, simply minimizing hydrophobic solvent-accessible surface area would predict. Such hydration forces might have an important influence on the protein folding pathway that a given polypeptide sequence would take, since we expect that variations in the solvation properties among different amino acids will be significant.<sup>33,35–37</sup>

If the water of hydration were to adopt a sufficiently large modification in structure relative to that of bulk water, it should result in a measurable difference of the wide angle scattering pattern in the region of the so-called “water ring”, i.e., the main diffraction peak at  $Q \approx 2.0 \text{ Å}^{-1}$  for water at room temperature. That the peak position of the first diffraction maximum is sensitive to water structure is well demonstrated by the fact that the peak shifts monotonically to lower  $Q$  as the temperature decreases, a trend that becomes even more pronounced in supercooled water.<sup>69–71</sup> Reducing the temperature leads to reduced distortion of hydrogen bonds and creation of an expanded configuration of molecules within the liquid structure, thereby causing a shift of the main peak to smaller angles, or larger effective Bragg spacings. We have reported our observation of a shift in the main diffraction peak for aqueous solutions of molecules with hydrophobic, but not hydrophilic, side chains, using neutron solution scattering experiments and molecular dynamics simulations.<sup>34,35</sup>

Neutron scattering experiments using both reactor (HFBR) and spallation (ISIS) sources were conducted on solutions of *N*-acetyl-L-amino acid-amide samples prepared as 1.0 mL of D<sub>2</sub>O added to 0.5 mmol dry reagent.<sup>34,35</sup> “Matched” solvent samples were prepared by the addition of sufficient H<sub>2</sub>O to imitate the hydrogen–deuterium exchange that occurs between the solute and the solvent.<sup>34,35</sup> Owing to the greater  $Q$  range of data collected at the ISIS spallation source ( $0.3 \text{ Å}^{-1} \leq Q \leq \sim 30.0 \text{ Å}^{-1}$ ), it is possible to put all measurements on an absolute scale.<sup>35,72</sup> An excess scattering intensity  $I_{\text{excess}}(Q)$ , where  $Q$  is the momentum transfer,  $Q = 4\pi \sin(\theta/2)/\lambda$ , was obtained by taking the difference between the scattering intensity measured for the solution and that measured for the matched solvent. The scattering measured for the matched solvent was scaled by  $k$ , the estimated number of water molecules per unit volume of a given solution divided by the number of water molecules per unit volume of pure water.

$$I_{\text{excess}}(Q) = I_{\text{solution}}(Q) - kI_{\text{pure water}}(Q) \quad (8)$$

Figure 4a portrays the HFBR and ISIS experimental excess scattering curves for the representative hydrophobic residue, NALA.<sup>34,35</sup> Together, the two experiments provide important confirmation that the effect seen for leucine is independent of the experimental setup. Figure 4b shows the HFBR excess scattering data for NALA and the hydrophilic amino acid, *N*-acetyl-glutamine-amide (NAQA), as well as the same simulated quantities for these same two amino acids (discussed further below). The concentration of NALA is 0.5 M (or  $\sim 1$  solute molecule per 100 waters), and the concentration of NAQA is 0.26 M; the NAQA curve was scaled by a factor of 2 to compare with the NALA results.

The difference in scattering between the NALA solution and pure water results in a shift of the main water diffraction peak to smaller  $Q$ , resulting in a ripple rather than a flat baseline, while the curve for NAQA appears flat in this region.<sup>34,35</sup> It is evident that the simulation data are in quite reasonable quantitative agreement in the region of the water ring ( $1.5 \text{ Å}^{-1} < Q < 2.5 \text{ Å}^{-1}$ ) with the neutron data for NALA, less so for NAQA, but exhibits the correct qualitative trends between NALA and NAQA. A set of control experiments to evaluate the scattering intensity for solutions of isobutanol (model NALA side chain) and *N*-acetyl-glycine-amide (NAGA, model NALA backbone) showed that the scattering for isobutanol also had a shift in the region of the water ring, while the main scattering peak for NAGA did not shift. This provided confirmation that the shift of the main water diffraction peak to smaller angle was due to the hydrophobic character of the NALA side chain and not the backbone. Note that the small angle agreement is poor, and we return to this point below when we discuss the simulations.

To interpret these experiments, molecular dynamics simulations were used to reproduce the measured scattering intensity and subsequently to analyze the molecular origin of the observed effect. The scattering intensity from an aqueous solution may be represented as a sum of intensities

$$I_{\text{solution}}(Q) = I_{\text{solute-solute}}(Q) + I_{\text{solute-water}}(Q) + I_{\text{water-water}}(Q) + I_{\text{intra}}(Q) \quad (9)$$

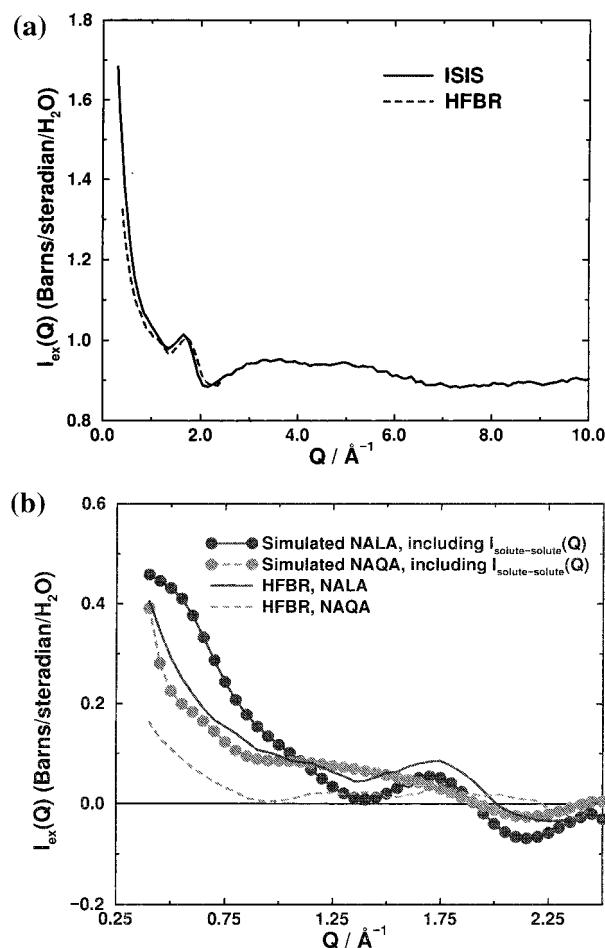
The first three terms in eq 9 arise from intermolecular correlations, and the last term refers to scattering interference between atoms on the same molecule. The neutron scattering contribution of each of the terms can be written as a sum of weighted structure factors,  $H(Q)$

$$I_{XY}(Q) = \sum_{\alpha}^n \sum_{\beta}^m c_X^{\alpha} c_Y^{\beta} b_X^{\alpha} b_Y^{\beta} H_{XY}^{\alpha\beta}(Q) \quad (10)$$

where

$$H_{XY}^{\alpha\beta}(Q) = 4\pi\rho \int_0^{\infty} r^2 [g_{XY}^{\alpha\beta}(r) - 1] \frac{\sin(Qr)}{Qr} dr \quad (11)$$

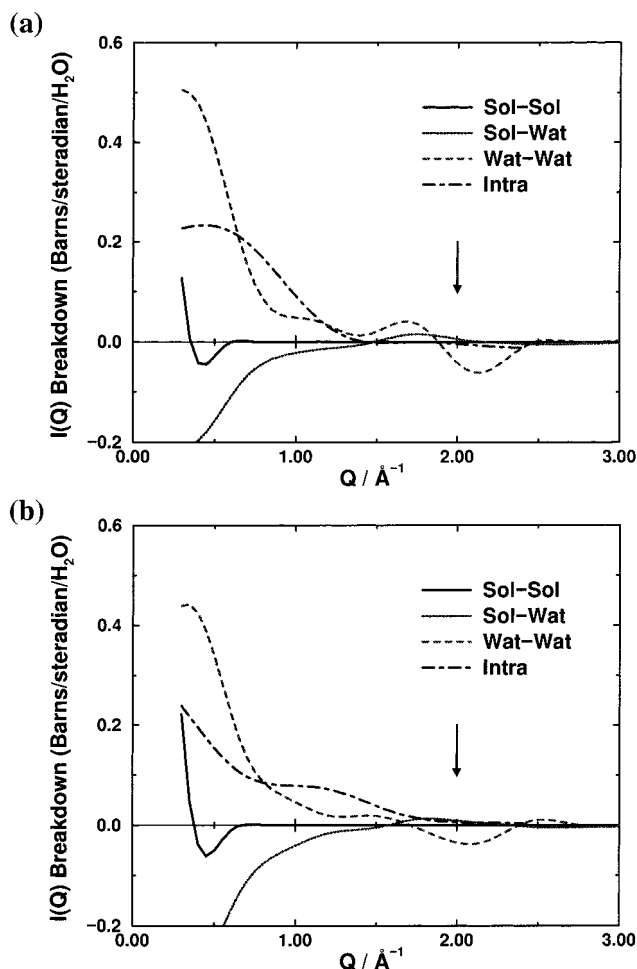
and  $X$  and  $Y$  correspond to solute or water, the indices  $\alpha$  and  $\beta$  refer to sums over atoms within a given molecule,  $\rho$  is the atomic density,  $g(r)$  is the radial distribution function,  $c$  is the atomic fraction, and  $b$  is the scattering length for an atom in the solute or solvent molecule. The experiments can be simulated by evaluating all radial distribution functions,  $g_{XY}^{\alpha\beta}(r)$ , and analyzing by grouping them into intramolecular and intermolecular contributions as in eq 9.



**Figure 4.** (a) Comparison of ISIS and HFBR experimental excess scattering curves,  $I_{\text{excess}}(Q)$ , in the region of the main water diffraction peak for NALA. The comparison shows that there is reasonable quantitative agreement between the HFBR and SANDALS experimental curves for NALA. (b) Comparison of the HFBR and simulated  $I_{\text{excess}}(Q)$  for NALA and NAQA. The comparison shows there is reasonable quantitative agreement between simulation and the HFBR data for NALA but only qualitative agreement between experiment and simulation for NAQA.

Simulation of the small-angle region is limited by both our simulation box size (which is valid for  $Q > 0.25 \text{\AA}^{-1}$ ) and adequate sampling over the full radial separation between all molecular centers in water. This is problematic for the solute–solute correlations, where the effective size of an individual solute is  $5.0\text{--}7.0 \text{\AA}$ . This accounts for the differences between experiment and simulation in the small-angle region, although in the next section we show how to interpret these differences in order to determine what are the leucine centers correlation in water.

A molecular dynamics simulation of a *single* solute in water was used to evaluate water–water  $g_{\text{OO}}(r)$ ,  $g_{\text{OH}}(r)$ , and  $g_{\text{HH}}(r)$  and solute–water  $g_{\text{OX}}(r)$  and  $g_{\text{HX}}(r)$  (where X is an atom of the solute) correlation functions to evaluate  $I_{\text{water–water}}(Q)$  and  $I_{\text{solute–water}}(Q)$ , eqs 9–11. The scattering length appropriate to deuterium is used to describe heavy water and all exchangeable hydrogens on the solute to match the experimental conditions. We use AMBER parameters<sup>73</sup> to describe a single amino acid in solution with enough SPC<sup>74</sup> water molecules to give the correct density. Further details of the simulation protocol is described elsewhere.<sup>36,75–77</sup>  $I_{\text{solute–water}}(Q)$  and  $I_{\text{water–water}}(Q)$  were multiplied by a factor of 4.56 and 2.60 for NALA and NAQA, respectively, to account for differences in the concentration of solute molecules between the simulated and experimental



**Figure 5.** Simulated individual contributions to  $I_{\text{excess}}(Q)$  for (a) NALA and (b) NAQA. The atomic simulations of solute–solute and solute–water correlations show that they are not significant contributors in the region  $1.5 \text{\AA}^{-1} < Q < 2.5 \text{\AA}^{-1}$  at the experimental concentrations. Intramolecular effects are also flat in this region. Essentially, the water–water correlations dominate the perturbation of the water ring, and this perturbation is attributable to alterations of the hydration shell around the hydrophobic amino acid.

conditions.<sup>34–38</sup> An independent simulation of pure water using 512 SPC water molecules was performed in order to generate the excess scattering differences.

To estimate the solute–solute correlations in water in the region of the water ring, we simulated 27 leucines and glutamines confined to a box  $30 \text{\AA}$  on edge, *without water*, to evaluate all radial distributions functions between all solute atomic pairs. This estimate takes into account the possible effects due to the smaller length scale and intersolute interactions, such as the formation of a solute–solute hydrogen bond, which might contribute to scattering in the water ring region. To obtain the contribution from intramolecular correlations to the scattered intensity for NALA and NAQA, the spherically averaged square of the molecular structure factor was calculated as

$$\langle F^2(Q) \rangle = \sum_a \sum_b b_a b_b \frac{\sin Q r_{ab}}{Q r_{ab}} \quad (12)$$

where  $r_{ab}$  is the distance between two atoms within one solute molecule.<sup>34</sup> The average was taken over a published library of molecular conformations for amino acids in proteins, weighting each one by its probability of occurrence.<sup>34,78</sup>

Figure 5 shows the individual contributions to  $I_{\text{excess}}(Q)$

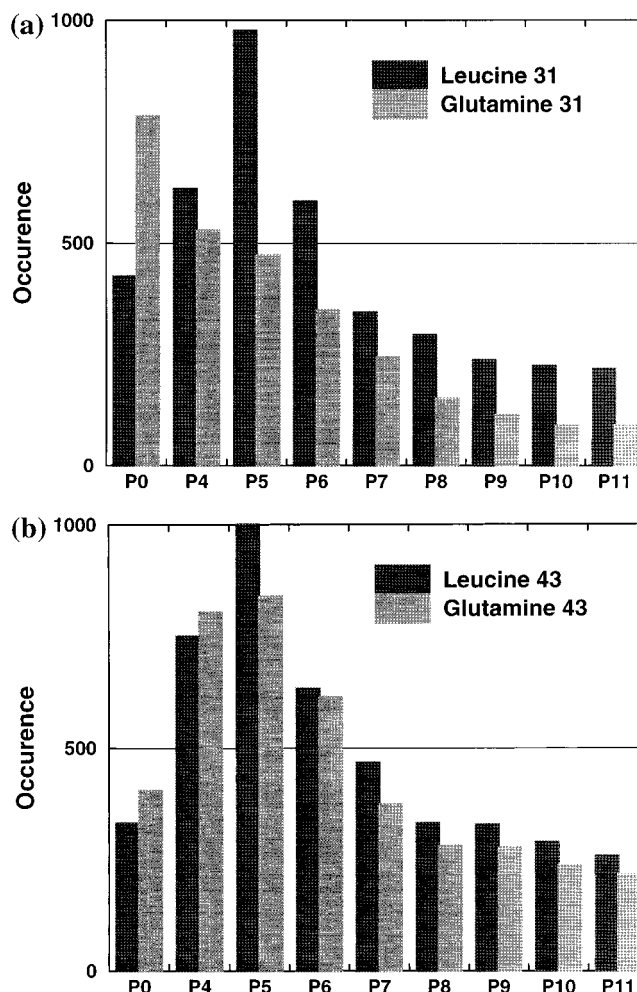
resulting from atomic simulations of solute–solute, solute–water, changes in water–water correlations between solution and pure water, and calculated intramolecular scattering. At the experimental concentrations ( $\sim 1$  NALA to 100 waters), there are no significant effects from either solute–water correlations or from hydrogen-bonding configurations between solutes. It is evident that the water–water correlations are the dominant contributor to the water ring signature for NALA, and therefore, the structural reorganization of water shifts the main water diffraction peak to smaller  $Q$ .<sup>36</sup> The curve for NAQA appears flat in this region and implies its hydration shell is largely equivalent in structure to that of the pure water background that has been subtracted.

We also used the MD simulations to provide insight into the molecular origin of the observed shift. A measure of hydration structure in terms of many-body functions is provided by the enumeration of non-short-circuited hydrogen bond pathways (defined by an energy or geometry criteria) in an associated liquid such as water.<sup>33,36,79,80</sup> The distribution function of water polygon sizes around the solutes is used to quantify differences in the organization of water around NALA and NAQA amino acids. Polygon distributions in different regions near the two solutes are displayed in Figure 6.  $P_N$  in these figures refers to the absolute number of polygons of each size: triangles ( $P_3$ ), quadrilaterals ( $P_4$ ), pentagons ( $P_5$ ), etc., counted over 2000 snapshots, except for  $P_0$ , which corresponds to the number of snapshots when no polygon of any size is found. The numbers next to the bar labels for NALA and NAQA indicate how many waters are present in any given region. Hydrogen bonds were defined by an energy cutoff,  $V_c$ , to be below  $-3.0$  kcal/mol.

A comparison of parts a and b of Figure 6 shows that structural differences between hydration shells around NALA and NAQA reside near the side chains and not their peptide backbones, consistent with the experimental observation of a shift of the main diffraction peak for isobutanol but not NAGA. In Figure 6a, the NALA side chain is seen to have a more ordered hydration shell, since  $P_0$  is nearly half that calculated for NAQA, and clearly, pentagons are an important feature of the polygon distribution around the NALA side chain. The idea that pentagons play a special role in enclosing the solute, similar to the role that pentagons play in fullerenes when compared to carbon sheets, gives topological justification for the importance and qualitative validity of clathrate analogies for describing hydrophobic hydration.<sup>33</sup>

The shift seen in the case of NALA is qualitatively analogous to the change in effective Bragg spacing observed in X-ray scattering when going from liquid water ( $\sim 3.1$  Å) to hexagonal ice ( $\sim 3.9$  Å). The expanded hydrogen-bonded network in ice is due to a dominance of hexagonal rings, which gives way in the liquid to both smaller and larger ring sizes, with greater distortion of these rings, so that the effective Bragg spacing in liquid water decreases. We find that dihedral angle distributions generated for pentagonal hydrogen-bonded rings show less puckering near the hydrophobic side chain. The greater number of more planar pentagons of hydration waters near the hydrophobic side chain of NALA is therefore consistent with the shift in the measured diffraction peak to a larger effective Bragg spacing, comparable to changes when transitioning from liquid to ice.<sup>36</sup>

The room-temperature scattering intensity difference between aqueous NALA solutions and pure water is similar to reported measured differences between ambient and supercooled water.<sup>69,70</sup> The measured shift of the main diffraction peak for the hydrophobic solutions is  $\sim 0.05$  Å<sup>-1</sup>, which corresponds roughly



**Figure 6.** Polygon distribution functions of water generated near the (a) side chain, and (b) backbone of NALA and NAQA. Hydrogen bonds were defined by an energy cutoff,  $V_c$ , to be below  $-3.0$  kcal/mol.  $P_0$  corresponds to the number of snapshots in which no polygon is found.  $P_5$  is the number of pentagons,  $P_6$  the number of hexagons, etc. The number in the legend corresponds to how many vertexes on average were used to generate the polygon distributions.

to a decrease of temperature to  $\sim 0$ – $10$  °C for the pure water liquid. Various structural interpretations of the shift in the main diffraction peak as water is supercooled have been put forward.<sup>81–85</sup> It has been suggested that the origin of the divergence in the temperature dependence at  $-46$  °C for supercooled water is due to a water network structure that favors the formation of larger, bulky polyhedra as the temperature is cooled,<sup>83,84</sup> or polyhedral faces that are pentagons in particular.<sup>81</sup> Similarly, hydrophobic groups may also be enclosed by similar polyhedral networks of water molecules. It has been hypothesized that the hydrophobic groups experience a water-induced mean attraction to maximize ideal hydrogen bonding sites between water ring pentagonal faces of the polyhedra, thereby providing a structural explanation for the thermodynamic driving force of hydrophobic attraction.<sup>83,84</sup> While we have made some definitive connection between the important role of pentagons in hydrophobic phenomena, the water structure connection between hydrophobic association and supercooled water remains unresolved.

### Evidence for Hydration Forces between Hydrophobic Amino Acids in Water

Alterations in the water structure around hydrophobic solutes is commonly thought to give rise to the unfavorable entropy at



room temperature that is a key signature of the hydrophobic effect.<sup>7</sup> There has been much research devoted to understanding hydrophobic phenomena in particular, and the consequences that hydrophobicity might play in the folding of real proteins have been made clear.<sup>1,19–23</sup> The question is still open, however, as to whether restructuring of the solvent within the hydration shell of hydrophobic residues results in significant thermodynamic forces between amino acids in a protein. The estimation of the range and magnitude of microscopic hydration forces, and its connection to water structure, requires the development of an approach that is sensitive to both water structure and any thermodynamic forces present due to hydration. We show in this section that the same combination of neutron scattering experiments and simulations<sup>34–38</sup> that revealed restructuring of solvent within the hydration shell of the NALA solute also provides evidence for longer-ranged hydration forces on the microscopic scale of the NALA amino acid.

Conceptually, we want to isolate  $I_{\text{solute-solute}}(Q)$  from the experimentally measured scattering function  $I_{\text{solution}}(Q)$  in eq 9. We then determine a model  $g_c(r)$  that best reproduces this excess signal,  $I_{\text{solute-solute}}(Q)$ . Once  $g_c(r)$  is determined, it can be related to hydration forces through

$$g_c(r) = e^{-W(r)/k_B T} \quad (13)$$

where  $W(r)$  is the “potential of mean force” between the two solutes separated by a distance  $r$ , i.e.,  $W(r)$  is a reduced free energy in which the explicit solvent configurations have been integrated out and all orientations and conformations of the two solute molecules have been spherically averaged. The importance of  $g_c(r)$ , which in turn defines  $W(r)$ , is that it describes the net correlations between solute pairs that implicitly account for the solvent environment.

In the solution scattering experiments we have reported,<sup>34–37</sup> the mole fraction of solute is quite small. We chose to work at these dilute concentrations, since we are trying to characterize hydration forces that are operative in early protein folding when the local concentration of amino acids is relatively dilute and residues are well-hydrated. However, the relative weight of all water–water contributions to the scattering intensity compared to solute–solute contributions is about 5000:1, which means the *direct* observation of solute–solute correlations is not possible because of the weak signal-to-noise ratio of solution scattering experiments.

Nonetheless, the scattering from water itself should allow the characterization of solute correlations in solution. That is because the solutes introduce new length scales into the water correlations that are due to their size, shape, and interactions. While the concentration of such correlated “holes” is still small, they are seen in the greater scattering contrast of water relative to the NALA solute (~36:1) after unwanted bulk water and water–solute correlations are removed. The solute–solute correlations are in fact directly related to the excluded volume effect seen in the water correlations. Our recent work gives formal reasoning on how to isolate solute–solute correlations in water and provides results for NALA correlations in solution.<sup>37</sup>

We start with an analysis of a recently proposed “uniform fluid” pair correlation function for the hydrogen–hydrogen (HH) correlations of water molecules that are excluded from a collection of *spherical* holes<sup>86</sup>

$$g_u^{\text{HH}}(r) = \left(1 - \frac{V_p}{V}\right)^{-2} \left[ 1 - 2\frac{V_p}{V} + \frac{V_p}{V} g_p^{\text{HH}}(r) + \left(\frac{V_p}{V}\right)^2 \frac{1}{v_p} \int d\mathbf{u} g_c(\mathbf{r}-\mathbf{u}) g_p^{\text{HH}}(\mathbf{u}) \right] \quad (14)$$

where  $V$  is the total system volume,  $V_p$  is the total volume occupied by the solutes,  $v_p$  is the volume occupied by an individual solute,  $g_c(r)$  is the radial distribution function for the solute centers, and  $g_p^{\text{HH}}(r)$  is the solute internal radial distribution function. Equation 14 can be further manipulated to isolate the solute centers pair correlation function,  $g_c(r)$ , which is especially important in the small-angle region. Assume initially that there is an ideal gas of solute molecules in solution so that  $g_c(r) = 1$  for all  $r$ . Equation 17 then reduces to

$$g_{\text{uncorr}}^{\text{HH}}(r) = \left(1 - \frac{V_p}{V}\right)^{-2} \left[ \left(1 - \frac{V_p}{V}\right)^2 + \frac{V_p}{V} g_p^{\text{HH}}(r) \right] \quad (15)$$

and when eq 15 is transformed to  $Q$  space, the result is

$$H_{\text{uncorr}}^{\text{HH}}(Q) = \left(1 - \frac{V_p}{V}\right)^{-2} \frac{V_p}{V} S_p^{\text{HH}}(Q) \quad (16)$$

The simulations discussed in the previous section describe the changes in the intermolecular pair correlations of water due to the presence of *one* solute, and they therefore include information about length scales in the water–water correlations due to *independent* or *uncorrelated* holes in water.<sup>86</sup> This is exactly the information contained in eqs 15 and 16. We can easily determine  $S_p^{\text{HH}}(Q)$  from a simulation of a single NALA in water and do not have to assume that NALA is spherical in shape.

We now manipulate  $g_u^{\text{HH}}(r)$  in eq 14 to separate the uncorrelated and correlated contributions and transform to  $Q$  space:

$$H_u^{\text{HH}}(Q) = H_{\text{uncorr}}^{\text{HH}}(Q) + \left(1 - \frac{V_p}{V}\right)^{-2} \left(\frac{V_p}{V}\right)^2 S_p^{\text{HH}}(Q) H_c(Q) \quad (17)$$

The more realistic estimates of the uncorrelated quantities, i.e., water correlations arising from a collection of uncorrelated NALA-shaped holes, can be subtracted from the experimental data to isolate an experimental signal that is due to the correlated quantities, the second term in eq 17.

Equivalently, we propose to isolate an experimental signal due to the correlated NALA solutes,  $I_{\text{solute-solute}}(Q)$ , by subtracting  $I_{\text{simulated}}(Q)$

$$I_{\text{simulated}}(Q) = I_{\text{solute-water}}(Q) + I_{\text{water-water}}(Q) + I_{\text{intra}}(Q) - kI_{\text{pure water}}(Q) \quad (18)$$

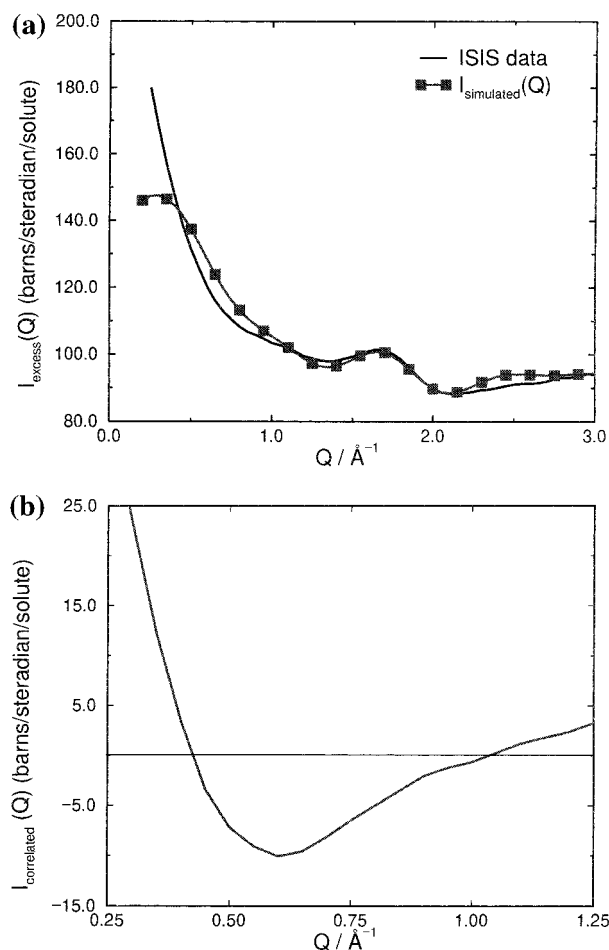
from  $I_{\text{excess}}(Q)$  obtained from the neutron scattering experiments. The remaining signal

$$I_{\text{correlated}}(Q) = I_{\text{excess}}(Q) - I_{\text{simulated}}(Q) \quad (19)$$

arises from the scattering of water molecules excluded from the solute regions where the solutes themselves are correlated in some way. Therefore, the intensity defined in eq 19 arises from the correlated term in eq 17, and model  $g_c(r)$ 's can be used to fit the remaining signal.

Figure 7a shows  $I_{\text{excess}}(Q)$  from the ISIS experiment along with  $I_{\text{simulated}}(Q)$ . The simulated results were offset by a factor equal to the sum of the squares of the scattering lengths for all NALA atoms, the theoretical limit for scattering at high  $Q$ . (In



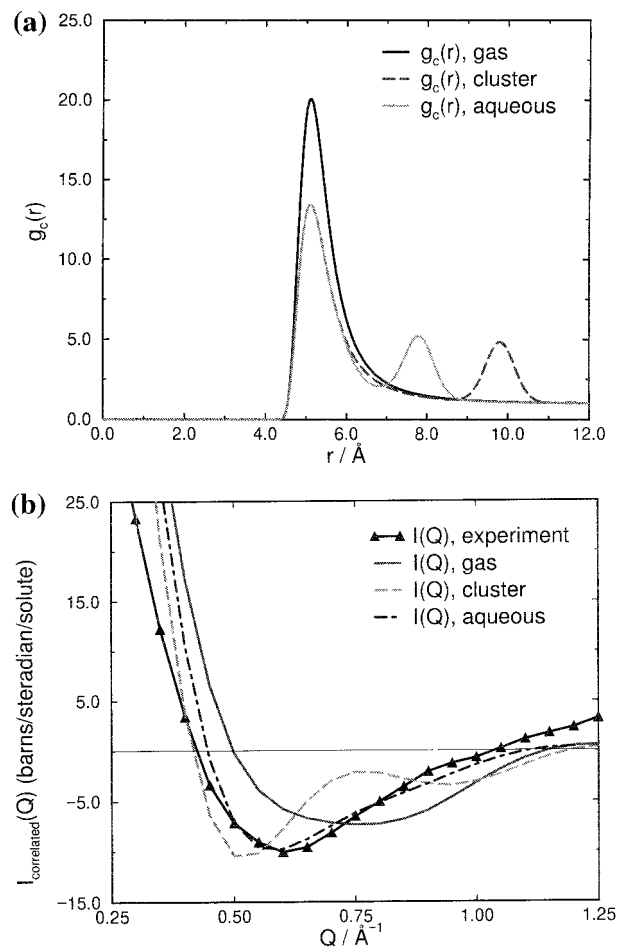


**Figure 7.** (a) ISIS neutron scattering data for a 0.5 M solution of NALA in D<sub>2</sub>O (solid line) and the simulated contributions from water–water, water–solute, and intramolecular scattering (squares). The data are in units of barns/steradian per solute molecule. (b) Experimental signal due to solute–solute correlations,  $I_{\text{correlated}}(Q)$ , that is obtained by subtracting the two curves in (a). The model  $g_c(r)$  that reproduces this curve describes the length scales of the NALA correlations in water.

ref 33, an incorrect scattering limit was used to put the simulated and integral equation results on the same scale as the experiment for NALA. The correct offset has been used in Figure 4a of the present paper.) Figure 7b exhibits the difference between the experimental curve and the simulated curve,  $I_{\text{correlated}}(Q)$ , which is nonzero over the range  $0.25 \text{ \AA}^{-1} < Q < 1.25 \text{ \AA}^{-1}$ . The importance of Figure 7b is that it represents the excess signal due solely to solute–solute correlations in aqueous solution. The next step is to determine the form of  $g_c(r)$  that reproduces  $I_{\text{correlated}}(Q)$  in Figure 7b.

In what follows we consider three qualitatively distinct solute centers pair correlation functions,  $g_c(r)$ : gas, cluster, and aqueous. While the entire space of solute–solute  $g_c(r)$ 's has not been exhaustively explored, we argue that all physically motivated  $g_c(r)$ 's have been considered with these three hypothetical functions; they in fact have qualitatively different peak positions. Once the qualitatively distinct solute–solute correlation functions are thus “enumerated”, further constraints on the values of peak positions and peak heights are imposed by the (approximately) known solute and water diameters, the density of solutes in solution, and constraints imposed by  $I_{\text{correlated}}(Q)$  itself.

Figure 8a shows the three qualitatively different examples of  $g_c(r)$ . The first is a gas of Lennard-Jones spheres. The second model  $g_c(r)$  is meant to exhibit ordering of NALA molecules



**Figure 8.** (a) Model  $g_c(r)$ 's describing gas, cluster, and aqueous forms of NALA correlations in water. The gas-phase  $g_c(r)$  corresponds to a Lennard-Jones interaction between solute molecules represented as spheres of  $\sigma = 5.0 \text{ \AA}$  with  $\epsilon = 1.6 \text{ kcal/mol}$ . The cluster  $g_c(r)$  is the same as the gas phase except for a second peak position at  $2\sigma$ . The aqueous model of  $g_c(r)$  is the same as the gas but has peak positions at  $\sigma_{\text{solute}} + \sigma_{\text{water}}$ . (b) Comparison of the excess experimental signal with the simulated solute–solute scattering derived from the various models of  $g_c(r)$  shown in (a). The comparison emphasizes that the gas and cluster forms of  $g_c(r)$  are probably not viable representations of the solute–solute correlations. The aqueous form is clearly in good agreement with the excess experimental signal.

as a cluster or liquid, with peak positions at  $\sigma$ ,  $2\sigma$ ,  $3\sigma$ , etc. The final form of  $g_c(r)$  that we consider is one that provides for positive correlations of NALA molecules at contact and separated by one or more water layers.<sup>87–92</sup> The presence of a solvent-separated minimum or minima implies that hydrophobic solutes in water are correlated over longer distances rather than arising from reducing exposed surface area (i.e., only being stabilized at contact). Figure 8b shows a comparison of the excess experimental signal due to solute–solute correlations and the simulated scattering for the gas, cluster, and aqueous models of  $g_c(r)$ . Neither the gas nor cluster forms reproduce the full range of experimental signal considered ( $0.25 \text{ \AA}^{-1} < Q < 1.25 \text{ \AA}^{-1}$ ).

While the agreement is not perfect, especially at the smallest angles considered, the aqueous form of  $g_c(r)$  is a much better description of  $I_{\text{correlated}}(Q)$  than either gas or cluster forms. While errors arising from the use of empirical force fields will always be an uncertainty, we have some confidence in the solute–water and water–water correlations obtained from simulation, since we have shown that the simulations can reproduce the NALA experiment in the region  $1.5 \text{ \AA}^{-1} < Q < 3.0 \text{ \AA}^{-1}$  with reasonable

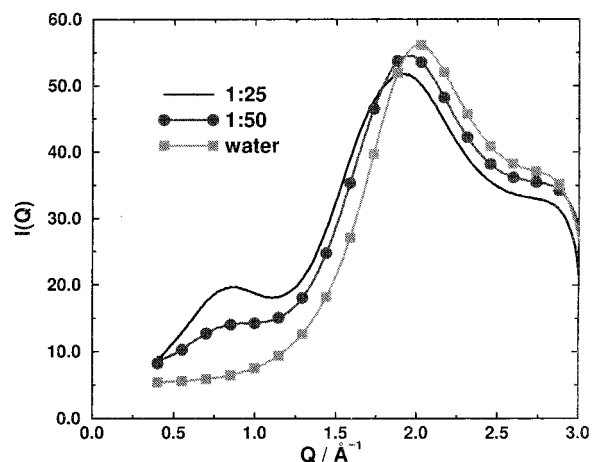
quantitative agreement.<sup>35,36</sup> Another possible source of error is the solute–water correlations, which may change from that calculated for a single solute in water, when the solutes themselves are in contact and/or solvent-separated. We have simulated the intensity contribution from solute–water correlations arising when two NALMA molecules are in contact and found no significant changes. Another potential error is that the intramolecular scattering is evaluated from a rotamer library based on globular proteins and may have different weights of side chain conformers than that exhibited in solution. We do not consider this to be in significant error, since protein surface residues, with side chains extending into solvent, would be strongly weighted in the library<sup>78</sup> of protein structures. We plan to do more careful experiments on a small-angle diffractometer in the future to help us better resolve  $I_{\text{excess}}(Q)$  for  $Q < 0.25 \text{ \AA}^{-1}$ .

### Aqueous Solvation for Concentrated Solutions of Hydrophobic Amino Acids

The solution scattering studies described in the previous two sections constitute a model of the solvation structure and free energy of amino acid association during early protein folding events. Solution scattering experiments and simulations can also be used to probe solvation forces for more concentrated aqueous solutions of hydrophobic solutes to mimic later folding stages when a significant fraction of the amino acids are sequestered into a hydrophobic core. In this section, aqueous solution X-ray scattering experiments and molecular dynamics simulations of concentrated solutions of NALMA are examined to characterize the solute distributions in water and to directly address what hydrophobic length scales are important in the later stages of protein folding.<sup>38</sup>

X-ray scattering intensities for pure water and aqueous solutions were measured with a rotating anode source and recorded using an R-AXIS-IV image plate camera routinely used for protein crystallography experiments. Ten minute exposures were required for the entire  $Q$  range of interest ( $0.3 \text{ \AA}^{-1} < Q < 3.0 \text{ \AA}^{-1}$ ). A circular integration code converts the raw detector image into a radial intensity curve, and the scattering curve is then corrected for absorption, for small variations in sample thickness, for the flat-plate detector geometry, and for polarization. Background scattering by air and the Kapton windows is obtained from measurements on an empty cell and subtracted from the experimental curves for each sample. Since the results of our pure water scattering experiments correspond very well with those of Nishikawa and Kitagawa,<sup>93</sup> we can use their curves to place our scattering on an absolute scale. We plan to report these data and more extensive experimental protocol in a future publication.<sup>38</sup>

Figure 9 displays the X-ray solution scattering measurements for NALMA in water at concentration ratios of solute to water of 1:50 and 1:25, along with the curve for pure water. The X-ray curve for the more dilute concentrations is dominated by the main X-ray diffraction peak of water at room temperature at  $Q \approx 2.0 \text{ \AA}^{-1}$ . However, at a concentration of 1:50, a new feature appears at  $Q \approx 0.8 \text{ \AA}^{-1}$  and develops into a peak at the saturated concentration of 1:25. Clearly, the new diffraction peak arises because of the presence of NALMA, but what is surprising is that the peak position shifts negligibly at the two measured concentrations, indicating that the new effective length scale represents a stable solute–solute configuration. In effect, the peak at  $Q \approx 0.8 \text{ \AA}^{-1}$  reflects the formation of a fluid, but ordered, phase, the amount of which depends on the total solute concentration but whose internal structure is not sensitive to solute concentration.



**Figure 9.** X-ray scattering intensity curves for pure water and NALMA in water solutions at concentrations of solute to water of 1:25 and 1:50. The data have been scaled to the pure water scattering data of Nishikawa and Kitagawa.<sup>93</sup>

We can use molecular dynamics simulations to interpret this new experimental feature at  $Q \approx 0.8 \text{ \AA}^{-1}$ . It is important to emphasize that we are unlikely to simulate the time progression involved in the formation of solute distributions seen experimentally, as this would require molecular dynamics simulations over very long time scales, and/or proper ensembles, to reach the final equilibrated distribution of solutes. For example, the simulation of the time progression of the formation of one large cluster might occur because of a single cooperative event, such as a large density fluctuation preceding a strong dewetting transition.<sup>21</sup> Such phenomena would be best simulated in the *NPT* ensemble, while our simulations were performed in the *NVT* ensemble. However, considerations of the mechanisms of how these solute configurations are reached are not important for this experiment. What is important is determining the final configurations of solutes that reproduce the static experimental observable.

We have focused therefore on what we believe is a representative diversity in the possible distributions of solutes seen experimentally. First, we consider a fully dispersed and hydrated configuration of NALMA molecules in water at concentrations of solute to water of 1:24 and 1:48. To prepare a dispersed configuration, a gas-phase simulation using the standard AMBER<sup>73</sup> energy function with 15 NALMAs in a box  $\sim 25 \text{ \AA}$  on edge was performed as described above but with all electrostatic interactions made repulsive by making all partial charges the same sign. Three separate gas-phase simulations were run for 10 ps each and then quenched, generating three uncorrelated snapshots of solute configurations in which the NALMAs were maximally dispersed. These configurations were overlaid on a configuration of pure SPC<sup>74</sup> water in the same size box, and waters overlapping the excluded volume of the solutes were deleted. These three maximally dispersed NALMA configurations in water were each equilibrated for 30 ps and radial distribution function statistics accumulated over an additional 30 ps. The simulations were kept short in order to realize an intensity curve that represented maximally dispersed solutes; the three sets of radial distribution functions were then averaged together to give the fully dispersed results.

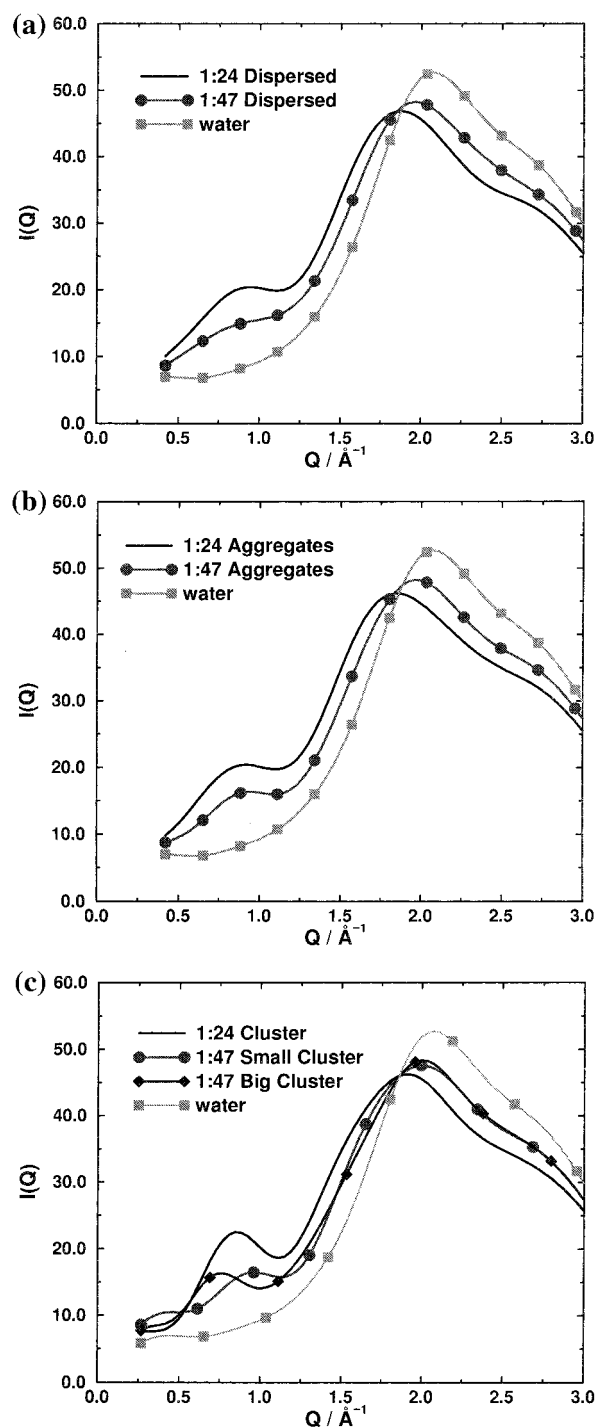
A second class of solute configuration is the formation of small molecular aggregates of solutes that range from mono-dispersed to clusters containing roughly four to six NALMAs in the most concentrated solutions. Concentration ratios of solute to water considered were in the range 1:24 to 1:100. For the

1:24 simulation, the solute preparation involved starting from a lattice configuration with 27 solutes in a box 25 Å on edge and deleting 12 sites at random to leave 15 in the box. This configuration was overlayed on a box of SPC water of the same size, deleting overlapping waters, and a very long equilibration phase of 400 ps was run during which NALMAs were found both isolated and aggregated into small clusters of sizes ranging from two to six. A subsequent 150 ps of statistics was run to collect  $g(r)$ 's and to generate the intensity curve for small molecular aggregates.

Finally, we consider the case that all NALMAs are configured into one cluster for concentrations of 1:24 and 1:47. The 1:24 solute configurations were generated by a gas-phase simulation of 15 NALMAs in a box 25 Å on edge, but in this case the solute-solute interactions were artificially enhanced by increasing the  $\epsilon$  parameter of the Lennard-Jones function for the  $C_\gamma$  carbon of the NALMA side chain. This gas-phase simulation was run for 20 ps and quenched at the end, generating a solute configuration in which the NALMAs formed a cylinder with a well-packed hydrophobic core. We also ran a similar gas-phase simulation with 18 NALMAs in a box 31 Å on edge for simulating a larger cluster at a concentration of 1:47. Each of these configurations was overlayed on a configuration of pure water in box sizes 25 and 31 Å on edge, respectively, and waters overlapping the excluded volume of the solutes were deleted. Short simulations (to preserve the character of the initial configuration) of 75 ps equilibration and 75 ps statistics were run to generate an X-ray intensity curve representative of fully clustered NALMA's. A simulation of a smaller cluster at a concentration of 1:47 was also run in a fashion similar to the 1:24 simulation.

Parts a, b, and c of Figure 10 shows the intensity curves derived from the simulations of the fully dispersed, small molecular aggregates, and single cluster simulations of NALMA solutes in water as a function of concentration, respectively. Figure 11 superimposes the simulations and experiment for the concentration of NALMA to water of 1:24, in which the observed feature at  $Q \approx 0.8 \text{ \AA}^{-1}$  is most developed. We find that the simulations do not distinguish between the maximally dispersed and small molecular aggregate configurations but that both resulting intensity curves are well differentiated from the intensity arising from a large NALMA cluster in water. More importantly, a comparison of the simulated data with Figure 9 shows that the concentration dependence seen experimentally is best reproduced by simulation for configurations of NALMA in which the solutes are maximally dispersed or involve small molecular clusters on the order of two to six NALMAs. When considering the single cluster data (Figure 10c), the scattering predicted for the smallest single cluster is too sharply defined and slightly shifted to a smaller  $Q$  value. This gets worse for the larger-sized cluster (which is simulated in a larger box and is therefore more dilute) where there is a significant shift to smaller  $Q$ .

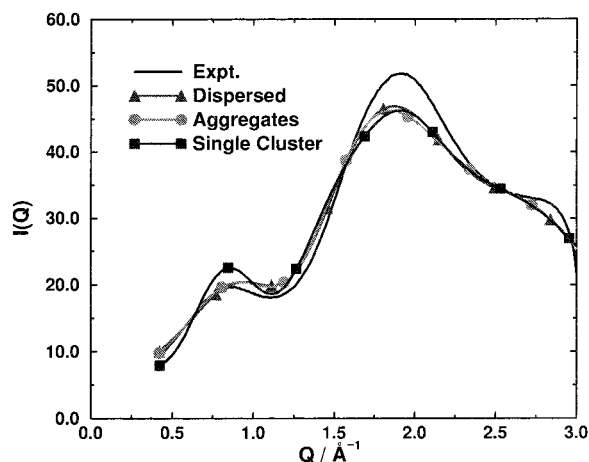
It is clear from Figures 10 and 11 that the simulations reproduce the same trends as a function of concentration as that from experiment when NALMA's are configured as fully dispersed or molecular aggregate configurations. However, we find that the simulated SPC X-ray scattering intensity for pure water does not adequately reproduce our experimental data curve. Preliminary analysis seems to suggest that SPC is "understructured" in that the position of the main water diffraction peak is at  $2.07 \text{ \AA}^{-1}$  instead of the experimental value of  $2.0 \text{ \AA}^{-1}$ . We find that other water models that have a more structured  $g_{OO}(r)$ , in particular, better reproduce our experimental



**Figure 10.** (a) Simulated X-ray scattering intensity curves for pure water and NALMA in water at concentrations of solute to water of 1:24 and 1:47, with NALMAs maximally dispersed. The data are calculated on an absolute scale. (b) Simulated X-ray scattering intensity curves for pure water and NALMA in water at concentrations of 1:24 and 1:47, with NALMAs configured as a distribution of small molecular aggregates. (c) Simulated X-ray scattering intensity curves for pure water and NALMA in water at concentrations of 1:24 and 1:47, with NALMAs configured as a single cluster.

scattering on pure water.<sup>94</sup> Similarly, our simulated solution scattering measurements tend to be "overstructured" in that the solute induces too large a shift of the main water diffraction peak to smaller angle, possibly overemphasizing a more open water network when hydrophobic groups are present. Further structural analysis of the simulations and experiment must be considered before such conclusions can be firmly made.<sup>95</sup>





**Figure 11.** Comparison of experiment and simulation for different solute configurations at the most concentrated value of solute to water of  $\sim 1:25$ . It is clear that the data are best reproduced by either fully dispersed or small clusters of NALMA, while the single cluster NALMA configurations are qualitatively inconsistent with experimental results.

Ultimately, existing protein and water force fields can be tested and modified when necessary, until the simulated neutron and X-ray scattering profiles quantitatively reproduce experimental results for a large variety of biologically relevant solutes in water.

Our experimental and simulated solution scattering experiments support a view that small hydrophobic domains are observed and therefore sustained in preference to large clusters for highly concentrated solutions of NALMA in water. This result should be extendible to real proteins, which are never purely hydrophobic. Large hydrophobic clusters would be observed if purely hydrophobic groups in water were considered at these high levels of concentration.<sup>21</sup> The fact that large clusters do not form emphasizes that protein folding and stability involves a detailed accounting of the complex hydrophilic and hydrophobic character of the protein backbone and side chains and balances a milieu of other interactions such as van der Waals, electrostatics, and amino acid side chain, and backbone conformational entropy costs, all of which compete on nanometer length scales.<sup>1,6,10,20,22</sup>

## Conclusions and Future Directions

It is widely appreciated that hydration forces are essential for protein stability, and they are also expected to play an important role in how quickly proteins fold to the correct native structure. The 36-mer lattice model examined in ref 39 is far from the complex reality of genuine proteins in aqueous solvent, but it possesses some of the essential features of protein folding, such as a unique ground state and a large set of possible conformations. By studying a lattice model that is closely related to many previous models with well-characterized kinetics and thermodynamics from over 20 years of studies, we have shown that incorporation of many-bodied solvation forces leads to faster folding, unique native states, and a more cooperative two-state folding transition. This lends support to the view that hydration forces are an important source of cooperativity in the protein folding transition. Our results indicate that the introduction of physically motivated solvation terms can improve the poor performance of two-flavor lattice models, since the multibodied nature of hydration mimics amino acid diversity, which in turn gives rise to a more cooperative folding transition, unique ground states, and faster folding.<sup>39</sup> We are currently extending

our folding studies to off-lattice simulations<sup>96</sup> and investigations that incorporate spatially long-ranged hydration forces into various protein folding models.

Our conclusion that a simple lattice model of protein folding can be improved with a more detailed description of solvation forces motivates further research into the experimental characterization of the solvation forces present between amino acid residues. In particular we have subtracted from a neutron solution scattering signal simulated quantities that describe uncorrelated NALA solutes in water, to leave an excess signal that contains information about the correlated solutes in water at dilute concentration. Various model pair distribution functions for NALA molecules, i.e., gas, cluster, and aqueous forms of  $g_c(r)$ , were tested for their ability to reproduce this excess experimental signal. We have found that the excess experimental signal is adequate enough to rule out gas and cluster pair correlation functions. The aqueous form of  $g_c(r)$  that exhibits a solvent-separated minimum, and possibly longer-ranged correlations as well, is not only physically sound but reproduces the experimental data reasonably well. The NALA scattering study at dilute concentration was designed to describe the nature of hydration biases in the earlier steps of folding when the local concentration of amino acids is relatively dilute.

We have also designed solution scattering studies to characterize the hydration of more concentrated amino acid solutions and to describe the consequences of hydration in later folding events when the local concentration of amino acids is high and driving toward the formation of a hydrophobic core. The analyzed X-ray solution scattering data are inconsistent with complete segregation of the hydrophobic solutes into one large cluster but instead show a distribution of monodispersed to small molecular aggregates of two to six hydrophobic amino acids being stabilized. Presumably the interactions that arise due to the complexity of the hydrophilic and hydrophobic character of the molecular protein sequence are equally important, complexity that is often ignored with an assumption that purely hydrophobic effects dominate.

More careful solution scattering experiments in the small-angle region are currently planned in order to resolve solute–solute correlations, their length scales, and thermodynamic consequences for amino acids other than NALA. Ultimately, the derived potentials of mean force, or “implicit” hydration potentials, could be interfaced with empirical protein force fields to be broadly used in computational studies of protein structure prediction and folding. Using the hydration potentials of mean force alone will clearly make exhaustive searches more feasible than with fully explicit models, while their greater complexity in comparison to lattice models might address important questions regarding more specific requirements for folding. During later stages of these simulated pathways we might usefully introduce the more detailed protein force fields to provide a tertiary structure prediction with atomic resolution.

Our analysis of solution scattering experiments on individual amino acids in water indicates that hydration structure around NALA is more ordered than water near NAQA, while the hydration water near their backbones is less ordered and largely equivalent between the two amino acids. The special role played by pentagons near the hydrophobic leucine side chain provides qualitative support for clathrate analogies, especially their topological role in enclosing hydrophobic solutes. The altered hydration structure near the leucine side chain, characterized by highly connected water vertexes forming rings that are dominated by planar pentagons in particular, extends roughly two solvation shells from the solute surface. This structural

persistence length suggests that *two* leucine peptides in water could sense each other's presence at a minimum distance of about 10 Å and could possibly entropically drive hydrophobic association over larger distances, since the two hydration shells confine water between them and further order the intervening bulk.

The characterization of the range and magnitude of hydration forces between individual amino acid side chains, and the connection to water structure, is a step toward defining the role of hydration in protein folding. We view the solution scattering experiments and simulations as a model systems approach similar to the determination of isolated secondary structure elements that might serve as folding intermediates but in the realm of hydration forces for solutes with full amino acid complexity. An especially important future direction for us is to extend our simulation and solution scattering experiments to polypeptide chains, to introduce the consequences of conformational entropy for model hydration intermediates in protein folding. Given our rather extensive understanding of NALA amino acids in water as a function of concentration, influences of the polypeptide backbone on biologically interesting sequences such as leucine zippers may further this goal.

**Acknowledgment.** J.M.S. is supported by a National Science Foundation Graduate Research fellowship. J.M.S. and T.H.G. thank the LDRD program through NERSC, U.S. Department of Energy Contract No. DE-AC-03-76SF00098, for support in FY98. G.H. and T.H.G. thank the Office of Biological and Environmental Research (OBER), U.S. Department of Energy Contract No. DE-AC03-76SF00098 for support in FY99. A.P. acknowledges support of NIH Grant GM-53163-01. T.H.G. gratefully acknowledges support from the Air Force Office of Sponsored Research, Grant No. FQ8671-9601129, and the National Energy Research Supercomputer Center for computer time. We thank Bob Glaeser for many interesting discussions and his careful readings of the manuscript, David Chandler for a preprint of ref 21, and Bing Jap for use of his X-ray machine to obtain our X-ray solution scattering data.

## References and Notes

- Dill, K. A. *Biochemistry* **1991**, 29, 7133.
- Onuchic, J.; Luthey-Schulten, Z.; Wolynes, P. *Annu. Rev. Phys. Chem.* **1997**, 48, 545.
- Baldwin R L. *J. Biomol. NMR* **1995**, 5, 103.
- Lazaridis, T.; Karplus, M. *Science* **1997**, 278, 1928.
- Dill, K. A.; Chan, H. *Nature Struct. Biol.* **1997**, 4, 10.
- Ben-Naim, A. *J. Chem. Phys.* **1989**, 90, 7412.
- Franks, F. *Water, A Comprehensive Treatise*; Plenum: New York, 1972–1982; Vols. 2–7.
- Alonso, D. O. V.; Dill, K. A. *Biochemistry* **1991**, 30, 5974.
- Stigter, D.; Alonso, D. O. V.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, 88, 4176.
- Rose, G. D.; Wolfenden, R. *Annu. Rev. Biophys. Biomol. Struct.* **1993**, 22, 381.
- Lewis, P. N.; Momany, F. A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1971**, 68, 2293.
- Anfinsen, C. B. *Science* **1973**, 181, 223.
- Kim, P. S.; Baldwin, R. L. *Annu. Rev. Biochem.* **1982**, 51, 459.
- Ptitsyn, O. B. *J. Protein Chem.* **1987**, 6, 272.
- Udgaonkar, J. B.; Baldwin, R. L. *Nature* **1988**, 335, 694.
- Roder, H.; Elove, G. A.; Englander, S. W. *Nature* **1988**, 335, 700.
- Kim, P. S.; Baldwin, R. L. *Annu. Rev. Biochem.* **1990**, 59, 631.
- Varlez, P.; Gronenborn, A. M.; Christensen, H.; et al. *Science* **1993**, 260, 1110.
- Hummer, G.; Garde, S.; Garcia, A. E.; Paulaitis, M. E.; et al. *J. Phys. Chem. B* **1998**, 102, 10469.
- Hummer, G.; Garde, S. *Phys. Rev. Lett.* **1998**, 80, 4193.
- Lum, K.; Chandler, D.; Weeks, J. D. *J. Phys. Chem. B*, in press.
- Rank, J. A.; Baker, D. *Protein Sci.* **1997**, 6, 347.
- Tsai, J.; Gerstein, M.; Levitt, M. *Protein Sci.* **1997**, 6, 2606.
- Gay, G.; Ruiz-Sanz, J.; Neira, J. L.; Itzhaki, L. S.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, 92, 3683.
- Fersht, A. R. *FEBS Lett.* **1993**, 325, 5.
- Sindelar, C. V.; Hendsch, Z. S.; Tidor, B. *Protein Sci.* **1998**, 7, 1898.
- Sharp, K. A. *Curr. Opin. Struct. Biol.* **1994**, 4, 234.
- Marqusee, S.; Baldwin, R. L. *Protein Folding: Deciphering the Second Half of the Genetic Code*; Gierasch, L. M., King, J., Eds.; American Association for the Advancement of Science: Washington, DC, 1990; p 85.
- Wright, P. E.; Dyson, H. J.; Walto, J. P.; Lerner, R. A. *Protein Folding: Deciphering the Second Half of the Genetic Code*; Gierasch, L. M., King, J., Eds.; American Association for the Advancement of Science: Washington, DC, 1990; p 85.
- Tobias, D. J.; Sneddon, S. F.; Brooks, C. L., III. *J. Mol. Biol.* **1990**, 216, 783.
- Tobias, D. J.; Brooks, C. L., III. *Biochemistry* **1991**, 30, 6059.
- Tobias, D. J.; Sneddon, S. F.; Brooks, C. L., III. *J. Mol. Biol.* **1992**, 227, 1244.
- Head-Gordon, T. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, 92, 8308.
- Pertsemidis, A. Ph.D. Thesis in Biophysics, University of California, Berkeley, 1995.
- Pertsemidis, A.; Saxena, A.; Soper, A. K.; Head-Gordon, T.; Glaeser, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, 93, 10769.
- Head-Gordon, T.; Sorenson, J. M.; Pertsemidis, A.; Glaeser, R. M. *Biophys. J.* **1997**, 73, 2106.
- Pertsemidis, A.; Soper, A. K.; Sorenson, J. M.; Head-Gordon, T. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, 96, 481.
- Hura, G.; Sorenson, J. M.; Glaeser, R. M.; Head-Gordon, T. *Perspect. Drug Discovery Des.*, in press.
- Sorenson, J. M.; Head-Gordon, T. *Fold Des.* **1998**, 3, 523.
- Sali, A.; Shakhnovich, E.; Karplus, M. *J. Mol. Biol.* **1994**, 235, 1614.
- Shakhnovich, E. *Curr. Opin. Struct. Biol.* **1997**, 7, 29.
- Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, 12, 183.
- Sfatos, C.; Gutin, A.; Shakhnovich, E. *Phys. Rev. E* **1993**, 48, 465.
- Gutin, A.; Shakhnovich, E. *J. Chem. Phys.* **1993**, 98, 8174.
- Shakhnovich, E.; Gutin, A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, 90, 7195.
- Pande, V.; Grosberg, A.; Tanaka, T. *J. Chem. Phys.* **1994**, 101, 8246.
- Socci, N.; Onuchic, J. *J. Chem. Phys.* **1995**, 103, 4732.
- Socci, N.; Onuchic, J. *J. Chem. Phys.* **1994**, 101, 1519.
- Yue, K.; Fiebig, K.; Thomas, P.; Chan, H.; Shakhnovich, E.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, 92, 325.
- Wolynes, P. *Nature Struct. Biol.* **1997**, 4, 871.
- Hao, M.-H.; Scheraga, H. *Physica A* **1997**, 244, 124.
- Hao, M.-H.; Scheraga, H. *J. Chem. Phys.* **1997**, 107, 8089.
- Hao, M.-H.; Scheraga, H. *J. Mol. Biol.* **1998**, 277, 973.
- Ferrenberg, A.; Swendsen, R. *Phys. Rev. Lett.* **1989**, 63, 1195.
- Onuchic, J.; Wolynes, P.; Luthey-Schulten, Z.; Socci, N. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, 92, 3626.
- Wolynes, P.; Onuchic, J.; Thirumalai, D. *Science* **1995**, 267, 1619.
- Socci, N.; Nymeyer, H.; Onuchic, J. *Physica D* **1997**, 107, 366.
- Marcelja, S.; Radi, N. *Chem. Phys. Lett.* **1976**, 42, 129.
- Parsegian, V. A. *Adv. Colloid Interface Sci.* **1982**, 16, 49.
- Israelachvili, J. N.; Pashley, R. M. *Nature* **1982**, 300, 341.
- Pashley, R. M.; McGuigan, P. M.; Ninham, B. W.; Evans, D. F. *Science* **1985**, 229, 1088.
- Rand, R. P.; Parsegian, V. A. *Biochim. Biophys. Acta* **1989**, 988, 351.
- Israelachvili, J. N.; Pashley, R. M. *Nature* **1983**, 306, 249.
- Israelachvili, J. N.; McGuigan, P. M. *Science* **1988**, 241, 795.
- Simon, S. A.; Fink, C. A.; Kenworthy, A. K.; McIntosh, T. J. *Biophys. J.* **1991**, 59, 538.
- Tsao, Y.-H.; Evans, D. F.; Wennerstrom, H. *Science* **1993**, 262, 548.
- Israelachvili, J.; Wennerstrom, H. *Nature* **1996**, 379, 213.
- Christensen, H. K.; Claesson, P. M. *Science* **1988**, 239, 390.
- Bosio, L.; Teixeira, J.; Dore, J. C.; Steytler, D.; Chieux, P. *Mol. Phys.* **1983**, 50, 733.
- Bellissent-Funel, M.-C.; Teixeira, J.; Bosio, L.; Dore, J.; Chieux, P. *Europhys. Lett.* **1986**, 2, 241.
- Dore, J. C. *J. Mol. Struct.* **1990**, 237, 221.
- Soper, A. K.; Howells, W. S.; Hannon, A. C. Report No. 89-046, Rutherford Appleton Laboratory, 1989.
- Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, 117, 5179.
- Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; p 331.

- (75) Swope, W. C.; Anderson, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1980**, *76*, 637.
- (76) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24.
- (77) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987.
- (78) Dunbrack, R. L., Jr.; Karplus, M. *Nature Struct. Biol.* **1994**, *1*, 334.
- (79) Rahman, A.; Stillinger, F. H. *J. Am. Chem. Soc.* **1973**, *95*, 7943.
- (80) Speedy, R. J.; Madura, J. D.; Jorgensen, W. L. *J. Phys. Chem.* **1987**, *91*, 909.
- (81) Speedy, R. J. *J. Phys. Chem.* **1984**, *88*, 3364.
- (82) Stanley, H. E.; Teixeira, J. *J. Chem. Phys.* **1980**, *73*, 3404.
- (83) Stillinger, F. H. *Science* **1980**, *209*, 451.
- (84) Stillinger, F. H. *Waters in Polymers*; Rowland, S. P., Ed.; ACS Symposium Series 127; American Chemical Society: Washington, DC, 1981; pp 11–22.
- (85) Walrafen, G. E.; Chu, Y. C. *J. Phys. Chem.* **1995**, *99*, 10635.
- (86) Soper, A. K. *J. Phys.: Condens. Matter* **1997**, *9*, 2399.
- (87) Geiger, A.; Rahman, A.; Stillinger, F. H. *J. Chem. Phys.* **1979**, *70*, 263.
- (88) Pratt, L. R.; Chandler, D. *J. Chem. Phys.* **1977**, *67*, 3683.
- (89) Pratt, L. R.; Chandler, D. *J. Chem. Phys.* **1980**, *73*, 3430.
- (90) Pratt, L. R.; Chandler, D. *J. Chem. Phys.* **1980**, *73*, 3434.
- (91) Pangali, C.; Rao, M.; Berne, B. J. *J. Chem. Phys.* **1982**, *81*, 2982.
- (92) Zichi, D. A.; Rossky, P. J. *J. Chem. Phys.* **1985**, *83*, 797.
- (93) Nishikawa, K.; Kitagawa, N. *Bull. Chem. Soc. Jpn.* **1980**, *53*, 2804.
- (94) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141.
- (95) Sorenson, J. M.; Hura, G.; Head-Gordon, T. In preparation.
- (96) Sorenson, J. M.; Head-Gordon, T. Submitted to *Proteins: Struct. Funct. Genet.*